

Міністерство освіти і науки України Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп'ютерний практикум №1

«Експериментальна оцінка ентропії
на символ джерела відкритого тексту»

Виконав

студент групи ФБ-93

Флекевчук Данило

Перевірила

Селюх П.В.

Київ – 2021

Мета:

Засвоїти поняття ентропії та надлишковості мови. Навчитись проводити частотний аналіз тексту та за його результатами рахувати ентропію тексту. Покращити навички програмування.

Завдання:

- уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- написати програму для підрахунку частот літер і частот біграм в тексті, а також підрахунку H_1 , H_2 за безпосереднім означенням. Підрахувати частоти літер та біграм, а також значення H_1 , H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 Мб), де ймовірності замінюються відповідними частотами. Також отримати значення H_1 , H_2 на тому ж самому тексті, в якому вилучено всі пробіли.
- за допомогою програми CoolPinkProgram оцінити значення $H(10)$, $H(20)$, $H(30)$.
- використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

I.

В цій частині роботи я рахую частоту входження літер, біграм(двох типів) у текстовий файл, та за цими показниками рахуємо ентропію мови. Для обчислення було використано мову програмування Python, так як вона найкраще підходить для важких обчислень. Для підрахунку входжень я використовував модуль collections а саме Counter для підрахунку входжень у текст та OrderdDict для сортування. Я розбив текст на різні типи біграм, з допомогою циклів з різними кроками (перехрестні: 2, прості: 1).

Код знаходиться в тій же папці що і протокол.

Далі я порахував ентропію для біграм та літер, і посортував літери за зростанням частоти. І вивів розраховані значення ентропії в консоль. Ентропію я рахував за формулою, що була наведена в завданні.

Для матриць з біграмами було виділено два файли, test_gap.txt та test.txt. Туди я повиводив значення частот для біграм у формі матриць, де на перехресті літер стоять біграми.

[illegible][illegible]

Літери.

З пробілами	Без пробілів
_ : 0.17018470956203863 а : 0.0660950216262635 б : 0.014432281039006059 в : 0.03839154172805269 г : 0.01401570956991704 д : 0.02657292659675882 е : 0.07232882162468782 ж : 0.009468876300924139 з : 0.012777812792978752 и : 0.053813549307092944 й : 0.00830877892365023 к : 0.027410008745046444 л : 0.03814534109620339 м : 0.02609234296338898 н : 0.05400164658982581 о : 0.0952116931512893 п : 0.02276961923595081 р : 0.034710349880641936 с : 0.0439221927219154 т : 0.053723932277099797 у : 0.02461316956723838 ф : 0.0010330578512396695 х : 0.007062018923965367 ц : 0.002300498703999874 ч : 0.015017253740279998 ш : 0.00683157513255442 щ : 0.002482687171568357 ы : 0.013700572761149934 ь : 0.019268646251053737 э : 0.0029258483088971 ю : 0.004661070362170978 я : 0.017726445493149712	а : 0.07965137392567485 б : 0.017392140247940356 в : 0.04626511057152555 г : 0.0168901357904256 д : 0.032022662593665964 е : 0.08716245244012143 ж : 0.011410810541381544 з : 0.01539836367908742 и : 0.06484995644547614 й : 0.01001279339965014 к : 0.033031418595764174 л : 0.04596841762973196 м : 0.03144351797128487 н : 0.06507662985300645 о : 0.11473828122218503 п : 0.027439350028838554 р : 0.041828957705827764 с : 0.052930020815976794 т : 0.06474196021466327 у : 0.02966098677698897 ф : 0.0012449235837659117 х : 0.00851034034240739 ц : 0.002772298848119323 ч : 0.018097082677641933 ш : 0.008232635748888589 щ : 0.0029918516250465807 ы : 0.016510368824929803 ь : 0.02322037639653368 э : 0.0035258989202750462 ю : 0.005616990774036282 я : 0.021361891809138618

Біграми (перші 10 в кожній категорії).

З пробілами		Без пробілів	
Перехресні	Прості	Перехресні	Прості
о_ : 0.011721	о_ : 0.011721	то: 0.018084	то: 0.009047
е_ : 0.009382	е_ : 0.009382	ов: 0.012589	ов: 0.006433
и_ : 0.008710	и_ : 0.008710	не: 0.012239	на: 0.006130
а_ : 0.008371	а_ : 0.008371	на: 0.012100	не: 0.006042
_в: 0.008361	_в: 0.008361	но: 0.011900	но: 0.005922
_п: 0.007913	_п: 0.007913	ст: 0.011619	ст: 0.005805
_с: 0.007845	_с: 0.007845	по: 0.010891	по: 0.005477
_н: 0.007757	_н: 0.007757	ко: 0.010675	ко: 0.005371
то: 0.007300	то: 0.007300	он: 0.010380	он: 0.005166
ь_ : 0.005971	ь_ : 0.005971	от: 0.009749	от: 0.004836

Без пробілів

Ентропія	Значення ентропії H	Надлишковість R
Літери	4.3497	13.00%
Біграми	3.9496	21.06%
Перехресні біграми	3.9502	21.00%

З пробілами

Ентропія	Значення ентропії H	Надлишковість R
Літери	2.1749	56.41%
Біграми	3.9496	21.06%
Перехресні біграми	3.9502	21.00%

Тут видно , що значення ентропії дуже сильно падає на літерах порівняно з іншими експериментами, бо пробіли дуже часто трапляюся у тексті.

II.

Для наступного завдання я використовую надану мені CoolPinkProgram. З допомогою цієї програми нам потрібно порахувати ентропію російської мови, а з нею і її надлишковість H(10), H(20), H(30).

Произвольная часть текста:
ы_о_нем_аналогично_
Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить Другой

Неравенство для энтропии:
2,52628436774245< H < 3.13669456263237
Двоичная таблица угаданных символов:
00100000000000000000000000000000
00100000000000000000000000000000
00000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0,42
q[2] = 0,1
q[3] = 0,08
q[4] = 0
q[5] = 0,02
q[6] = 0,02
q[7] = 0,06
q[8] = 0
q[9] = 0,02
q[10] = 0,02
q[11] = 0
q[12] = 0,02
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0,04
q[18] = 0,02
q[19] = 0
q[20] = 0
q[21] = 0,04
q[22] = 0
q[23] = 0,02
q[24] = 0
q[25] = 0,06
q[26] = 0
q[27] = 0
q[28] = 0,02
q[29] = 0
q[30] = 0,02
q[31] = 0,02
q[32] = 0

Строка состояния:

Ентропія	Значення ентропії	Надлишковість
$H(10)$	3.0000	40.00%
$H(20)$	2.8140	43.72%
$H(30)$	1.9460	61.08%
H	2,5926	48.00%

Надлишковість рахується за формулою $1 - (H/H(0))$, де $H(0)$ для російської мови рівне 5. У мене це вийшло 0,48, я думаю такт мала надлишковість зумовлена моїми неякісними знаннями російської мови.

Висновок:

В ході роботи я навчився обраховувати ентропію джерела тексту для символів та біграм, також я навився обраховувати надлишковість тексту. Для російської мови надлишковість склала 48%, за вказаними вище причинами. Для пошуку ентропії біграм та літер було обрано “Злочин і кара”. Якщо не забрати пробіли то, тоді вони будуть найбільш вживаним символом (“_” : 172811), якщо ні то це літери: “о” : 96681, “е” : 73380 та “а” : 67116 відповідно, а най менш вживані: “ц” : 2336, “ф” : 1049. Найчастішими біграмами з пробілом виявились: “о_” : 0.023578, “е_” : 0.018882, “и_” : 0.017332, а без пробілів: “то” : 0.018084, “ов” : 0.012589, “не” : 0.012239. Наявність пробілів дуже полегшує роботу криптоаналітика.